Mining for putative regulatory elements in the yeast genome using gene expression data

Jaak Vilo[*], Alvis Brazma[*], Inge Jonassen[†], Alan Robinson[*], Esko Ukkonen[‡]

Abstract

We have developed a method for discovering signals in large sets of genomic sequences. The method is based on exhaustive search for *a priori* unknown statistically significant sequence patterns of unrestricted length and their subsequent clustering by similarity. The statistical significance is measured taking into account the background distribution, i.e., the occurrence of the patterns in a superset of sequences, which makes the method sufficiently sensitive for unsupervised discovery of transcription factor binding sites using gene expression data. The clustering of the discovered patterns by similarity makes it possible to report thousands of statistically significant pattern in an accessible way for a human investigator.

We demonstrate the method by performing a systematic analysis of upstream regions of all yeast genes. We use public gene expression data to cluster the genes according to their expression profiles, assessing the "goodness" of each cluster by a standard measure taken from the literature. For each cluster we retrieve sequences upstream to the respective genes and exhaustively search for all the sequence patterns of unrestricted length that are present in the clusters in at least 10 of sequences. For each pattern we calculate a measure of significance based on the pattern's occurrence frequency in upstream regions of genes in the cluster versus it occurrence frequency in all upstream regions. We repeat the same experiment for randomized data to assess the significance threshold for such patterns. We find that there are more than 6000 statistically significant substring patterns. Several of the clusters have highly homologous upstream sequences. We concentrate on studies of the patterns discovered from less homologous sequences. There are about 1500 such significant patterns. We use our pattern grouping method, to group them in 62 groups of similar patterns, and for each group we find an approximate IUPAC consensus?***. We match all 1500 patterns against the experimentally proven yeast binding sites as given in SCPD database. Most of the clusters have strong matches in the database *** ...

1 Introduction

A collection of gene expression level measurements taken under various experimental conditions by microarray or any other technology, define expression profiles of the respective genes. It

[*][vilo,brazma,alan]@ebi.ac.uk; European Bioinformatics Institute EBI ; EMBL Outstation; Wellcome Trust Genome Campus; Hinxton, Cambridge CB10 1SD ; United Kingdom

[†]inge@ii.uib.no; Department of Informatics; University of Bergen; HIB; N5020 Bergen, Norway

[‡]ukkonen@cs.helsinki.fi; Department of Computer Science, P.O.Box 26, FIN-00014 University of Helsinki, Finland;

seems reasonable to hypothesize that genes with similar expression profiles, i.e., genes that are *coexpressed*, should share something common in their regulatory mechanisms, i.e. should be *coregulated*. Therefore by clustering together genes with similar expression profiles we can find groups of potentially coregulated genes.

Evidently a major role in gene regulation in eukaryotic organisms is played by specific proteins, called *transcription factors*. By binding to sequence specific sites in the DNA, called *transcription factor binding sites*, they facilitate or hinder the transcription of a particular gene. The transcription factor binding sites are located in *promoter regions*. In yeast these regions are predominantly (but not exclusively) in the immediate vicinity of the gene (typically less than a thousand bp upstream of the translation start site).

During the last couple of years microarray based gene expression data have been widely used to find potentially coregulated genes and binding sites in their promoters. De Risi et al [7] have studied the diauxic shift in yeast. They found several distinct clusters in the gene expression profiles, and were able to show the presence of several transcription factor binding sites located upstream to the respective genes. The dataset by De Risi at al [7], was also studied by Van Helden et al [24], where they systematically searched for oligonucleotides overrepresented upstream to potentially coregulated genes and showed that potential transcription factor binding sites can be found in this way. A systematic analysis of overrepresented sequence patterns in clusters of upstream sequences obtained by clustering the diauxic shift expression profiles has been done in [4]. It was shown that overrepresented patterns present in sequences upstream from genes from expression profile clusters of at least 25 genes cannot be explained by statistical "chance". Many of the discovered patterns have matches in known yeast transcription binding site descriptions.

In a more recent study, Chu et al [6] and Spellman et al [20] studied yeast gene expression profiles during sporulation and cell cycle respectively. Tavazoie et al [22] clustered 3000 expression profiles of the most variable yeast genes during the cell cycle (15 time points, data from Cho et al [5]) into 30 clusters by using K-means algorithm. They found that for a half of these clusters, strong sequence patterns are present in upstream sequences. Moreover, it was noted that upstream sequences from genes from the tightest expression profile clusters (i.e., the clusters with smaller average Euclidean distance to their centers) contain more significant patterns. In a different study Wolfsberg et al [25] studied the discovery of potential regulatory sequences from groups of genes whose expression was peaked at distinctive phases of cell cycle as identified by Cho et al [5].

The result of the expression profile clustering depends on the choice of the distance measure in the expression profile space, as well as on the clustering algorithm itself. Apparently there is no objectively "right way" of clustering the expression profiles, since various elements in each profile may be influenced by some particular regulation aspects. Moreover, there are indications that in eukaryotic organisms genes are not regulated by on/off switches, but more continuously. An alternative to fixing a particular clustering is to study a number of different clusterings in parallel, looking for overrepresented sequence patterns obtained from each of them.

Note that the promoter sequence analysis is an algorithmically hard and nontrivial task due to many factors. First, coexpression does not necessarily mean coregulation: several regulation mechanisms may lead to the same expression patterns for the given experimental conditions. Next, currently the gene expression measurements are not very accurate, which means that the expression profile clustering is inherently fuzzy. Finally, the identification of promoter regions is currently unreliable, therefore even for truly coregulated genes the sets of promoter regions may be inaccurate. As a result, usually we have to deal with sets of potential promoter regions of potentially coregulated genes. Many of the sequences or large parts of their contents may actually not be related to transcription regulation. Also, it is not known beforehand which sequences share similar regulatory mechanisms. Therefore, the algorithms used for transcription factor binding site prediction may have to detect only marginally overrepresented patterns in sets of up to hundreds of sequences of length of thousands of nucleotides.

We have developed a method for discovering signals in large sets of genome sequences, which is based on exhaustive search for *a priori* unknown statistically significant sequence patterns of unrestricted length and their subsequent clustering by similarity. The statistical significance is measured taking into account the background distribution, i.e., the occurrence of the patterns in a superset of sequences. The clustering of the discovered patterns by similarity makes it possible to report thousands of statistically significant pattern in an accessible way for a human investigator. We have implemented the method as a Web based tool.

To demonstrate the method, we clustered systematically all yeast genes by expression profiles comprising 80 experimental conditions taken from [**?**] by k-means clustering algorithm. For each cluster we measure its so-called average silhouette value, which characterize its "goodness". Choosing different k values and the initial partitioning, we obtaining over 52 thousand different clusters (many of these clusters are highly overlapping). For each of these clusters we retrieve sequences 600 bp upstream of the respective gene, and exhaustively searched all the sequence patterns of unrestricted length that are overrepresented in the sequences of the cluster. For each pattern we measure its significance, based on its occurrence frequency in all the upstream sequences assuming a binomial distribution. We repeat the same for randomized data to assess the significance threshold for such patterns. We find that several of the gene clusters contain highly homologous upstream sequences. Notably, the expression profiles of these clusters have the highest "goodness" values. We concentrate on studies of the patterns discovered from less homologous sequences, and there are about 1500 such significant patterns. We use our pattern grouping method, to group them in 62 groups of similar patterns, and ***for each group we find an approximate IUPAC consensus?***. We match all 1500 patterns against the experimentally proven yeast binding sites as given in SCPD database. Most of the clusters have strong matches in the database ***

The aims of these studies is to mine for new, statistically significant patterns in the putative regulatory regions, which can become candidates of further studies by more conventional means. Such data mining experiments are not a substitute for "conventional single-gene dissections" [**?**], on the contrary their aim is to explore simultaneously thousands of genes (which cannot be done by conventional methods) to generate targets for conventional studies.

2 Clustering the gene expression profiles

A number of approaches have been used for gene expression profile clustering by different groups. The classic hierarchical [10, 9] and partitioning based (K-means) clustering algorithms [12, 22], as well as self-organizing maps [15, 21] have been all used for clustering the expression profiles. Even a simple clustering algorithm based on binning (i.e., discretizing the expression profile space and clustering together the profiles that map into the same pattern of bins) has shown to be useful for discovering transcription factor binding sites [4]. More recently new algorithms have been developed specifically for gene expression profile clustering [2] based on finding approximate clusters in graphs.

One of the most important questions in clustering the expression profiles is how to define the profile similarity measure. It has been generally acknowledged that currently we do not know what is the most appropriate way to measure such similarities. In absence of any theory, Euclidean distance and linear correlation coefficient based similarity measures are the most widely used. The usage of various similarity measures for gene expression profile clustering has been discussed in [8], where measures based on rank correlation and mutual information have been proposed.

We have implemented a tool that combines facilities of gene expression profile clustering and sequence pattern discovery as a Web-based service (the tool will be published elsewhere). Currently, our own implementation can perform the hierarchical (average linkage, maximal linkage, and single linkage) and partitioning-based (K-means) clustering algorithms for a number of profile similarity measures (see http://www.ebi.ac.uk/microarray/???). The performance of our implementation is noteworthy: clustering of the 6221 genes with 80 individual experiment measurements for each gene uses less than 80MB memory and can be run in less than 2 minutes (distance matrix calculation takes about 50-60 seconds, and hierarchical clustering about 15-20 seconds) on a 500MHz Intel Pentium III workstation running Linux operating system. The web interface allowed us also to create an interactive visualization of the clustering in the form of graphical imagemaps.

For many applications it is important to be able to evaluate the "goodness" of a cluster, i.e, how close are its elements, and how far are they from next closest cluster. Such a "goodness" measure has been proposed in [19] and is based on the notion of *silhouette plot* and an average silhouette value of a cluster (for a detailed definition see [19]).

The silhouette value $s(i)$ lies between -1 and 1. If $s(i) = 1$, the object is well classified, if $s(i) < 0$, the object is badly classified (closer to members of some other cluster than A). The average silhouette value for a cluster can be used as a measure of the "goodness" of that cluster. The silhouette value characterizes not only the "tightness" of the given cluster, but also how far are each element of the cluster from the next closest cluster.

3 Sequence pattern discovery

By sequence pattern discovery we mean finding of *a priori* unknown patterns (of some given class, such regular expressions or oligonucleotides) that are statistically overrepresented in given set of sequences (given some background distribution). Algorithms for sequence pattern discovery have been widely used for characterizing protein families (e.g., [14], for survey see [3]). Relatively more recently several sequence pattern discovery systems have been applied to regulatory element analysis [1, 11]. In [18] a new algorithm based on Gibbs Motif Sampling [17], has been developed, specifically for discovering transcription factor binding sites.

A suffix-tree ([16, 23]) based sequence pattern discovery algorithm, which allows one to enumerate and produce systematically and exhaustively different kinds of sequence patterns of unrestricted length, has been developed in [**?**]. The algorithm has been used in [4] to discover regular expression type patterns in gene upstream sequences. A *substring pattern* is a word over a DNA alphabet $\{\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}\}$, and we say that a pattern π matches a sequence α, if there exist (possibly empty) strings α' and α'' over the DNA alphabet, such that $\alpha = \alpha'\pi\alpha''$.

Given a set S of N sequences, a subset $C \subseteq S$ of size n, and a pattern π that occurs in k sequences from C, we can ask how probable is the event of observing the pattern π to match k or more sequences of C.

We calculate that probability according to a binomial distribution. We estimate the expected probability p that pattern π matches individual sequence of C from the total number of occurrences K of that particular pattern in the set of all N sequences, $p = K/N$. We call p the *background probability*. Given k occurrences in the set C, we ask how probable this is given the background distribution p? According to the binomial distribution, the probability of a pattern occurring in exactly k sequences "by chance" is:

$$P(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

The probability of a pattern π to occur k or more times in a set of C sequences is:

$$P(\pi, C) = \sum_{i=k}^{n} P(i, n, p)$$

We will use the inverse of this probability as the pattern score for a pattern π in respect to a particular set of sequences C. The smaller the probability, the higher is the pattern score. We have used this probability to score all patterns in all sets of upstream sequences of clusters of coexpressed genes.

4 Grouping patterns by similarity

One of the problems with the approach in [4] is that typically there are too many significant regular expression type patterns in a cluster of coexpressed genes for a human to study each of them in detail. On the other hand, it seemed that many of these patterns fall into groups of similar patterns, and possibly each such group is related to the same regulatory element. Moreover, clustering the discovered patterns by similarity makes it considerably simpler to study them. A cluster can presented as an alignment of similar patterns, and a profile [**?**] can be easily derived from such an alignment.

To group patterns by similarity, we need to define a similarity or a distance measure for a pair of patterns π_1, π_2. The information content of a pattern is defined as in [**?**]. A pattern π_{12} is found that is a generalization both of π_1 and of π_2, and π_{12} is chosen so that it has the maximum information content over all patterns having this property and being within a defined pattern class. Then, the distance between pattern π_1 and π_2 is defined as

$$d(\pi_1, \pi_2) = \frac{I(\pi_{12})}{\min(I(\pi_1), I(\pi_2))}, \tag{1}$$

where $I(\pi)$ is the information content of pattern π. In our experiments we used substring patterns, and we restricted also patterns used for measuring syntactic distance, to be substring patterns, which makes the calculation of the distance measure straight forward.

Having defined the similarity measure between patterns, we use standard hierarchical clustering algorithm to cluster them.

We use the pattern discovery algorithm SPEXS to find the longest substring that is common to all (or given percentage) of the patterns in each cluster. We use this substring as an anchor to align the patterns of the cluster (for instance see Figure 3).

5 Experiment

We performed a large-scale experiment of yeast expression and sequence data analysis. We used the data[1] combining various yeast expression experiments performed in Stanford University [9] and performed a following computational experiment. Note that the below described steps have been performed in a highly automated way.

1. **Clustering the expression data.** We clustered 6221 genes based on their expression profiles by the K-means clustering algorithm using Euclidean distance in 80-dimensional space. We varied the value of K (the number of clusters) between 2 and 1000 and repeated the clustering for each selected K ten times with different random sets of initial cluster centers. In total we did over 900 separate clusterings. For each cluster we computed the average silhouette (Section 2) value. We selected the clusters of size between 20 to 100 genes and obtained in this way over 52,100 different clusters.

2. **Sequence pattern discovery.** For each cluster we took the set of gene upstream sequences of length 600 bp and enumerated all patterns occurring in at least 10 of these. We scored all patterns according to the probability of their occurrences in the cluster using binomial distribution and background probability estimation as described in Section 3.

3. **Controll experiment.** To assess the statistical significance threshold for the patterns we repeated the same experiment on the randomized data by taking exactly the same clusters but shuffling the gene names so that the clusters consisted of random sets of upstream sequences.

4. **Finding the significance threshold.** We plotted the average silhouette value and the score of the best pattern of each cluster on a two-dimensional plot in Figure 1 (top left), and for the randomized data similarly Figure 2 (top right).

5. **Pattern selection.** Totally there are over 6000 significant patterns (see Table 1 for sample patterns). We study the distribution how many significant patterns are found in each gene cluster. The distribution suggest that the clusters can be divided into two groups: the ones producing more than 600 patterns [2], and the ones producing less then 600 patterns. We study each group separately.

6. **Removing homologous sequences.** We studied the upstream sequences of the clusters that produce more than 600 patterns by Clustal W and found that the sequences in these clusters are highly homologous. ***Inge – is the a number that characterizes the homology – the alignments show high similarity, but is there a way to quantify this?*** There are such 508 clusters, but they contain only different 169 ORFs (see Figure 1 bottom).

[1]The data set consists of gene expression levels of 6221 yeast genes with a total of 80 experimental conditions. These 80 measurements are related to time course analysis of yeast cell cultures during the cell cycle, sporulation, and diauxic shift experiments. The data has been downloaded from P. Brown's laboratory (http://rana.stanford.edu/).

[2]The reason for cutoff at 600 "good" patterns produced from one cluster was set based on the jump in number of patterns from up to 218 it was more or less continuous, then the next values were 336 and next 746.

7. **Interesting patterns.** The homology of sequences in the clusters that produce smaller number of patterns are low, therefore the significant patterns in these clusters are candidates for regulatory signals in the sequences. These clusters together contain 3727 different ORFs. There are about 1500 such patterns, which is still too many to study them one by one.

8. **Grouping the patterns.** We group these patterns based on common information content based similarity measure and produce 62 clusters of similar patterns. For each cluster find an approximate IUPAC consensus pattern by SPEXS.

9. **Evaluation of discovered patterns against known TF binding sites.** We systematically explore all over 1500 patterns produced my nonhomologous upstream sequences by matching them against all experimentally proved yeast binding sites as given in SCPD [3]. We regarded a pattern and a site as matching if either of them is completely contained the other. We regarded a cluster of patterns as matching the site, if one of the patterns in the cluster were matching it. 48 of the 62 pattern clusters had matches in SCPD. Some of the sites in SCPD are rather short and can have matches by chance (in fact there are many sites consisting of one nucleotide, and we excluded them before matching). The results are summarized in Table 2

6 Results

Table 2 show consensus pattern, used in here for naming of the clusters only, has been calculated from alignments. The nucleotide groups (IUPAC groups, represented in here using a regular expression notation) have been introduced when the frequency of the less frequent nucleotide in respective is over 25% of the frequency of the more frequent nucleotide. Inside the groups nucleotides are ordered based on their frequency. Lowercase is used when the majority of the patterns does not have any nucleotide in that position. Or in other words, when the most frequent nucleotide in the respective column should be a dash.

Our observation that the pattern and cluster scores correlate is consistent with the observation of Tavazoie et al [22]. However, we have done a more systematic experiment for precisely defined pattern and cluster "goodness" measures for considerably more clusters reporting also the numeric evidence. Although the observation is not surprising, it suggests that the fast and simple K-means clustering algorithm can be used in the large scale analysis of all genes of an organism. It enables the finding of coexpressed genes based on the expression profiles allowing the consecutive search for coregulated genes.

We studied more closely some of the clusters and the respective discovered patterns. For instance, the expression profile for the cluster that produced the 15th highest scoring pattern `GGTGGCAA` is shown in Table 3 (we will refer to this cluster as `GGTGGCAA`-cluster). This pattern occurs in 20 out of 40 upstream sequences of the cluster, and only in 96 out of all over 6200 upstreams. Note that most of the genes are related to functioning of the proteasome.

We also studied how our discovered patterns compare to experimentally proven binding sites in yeast by comparing them to SCPD database. For instance, two of the pattern clusters (cluster nr. 15 and 42 in the numeration produced by the clustering algorithm) had matches

[3] To be able to perform such high-throughput matching of more than 1500 patterns against all experimentally proven sites in SCPD, we had to download the SCPD locally using CGI script. ***Jaak- how bloody hard it was***.

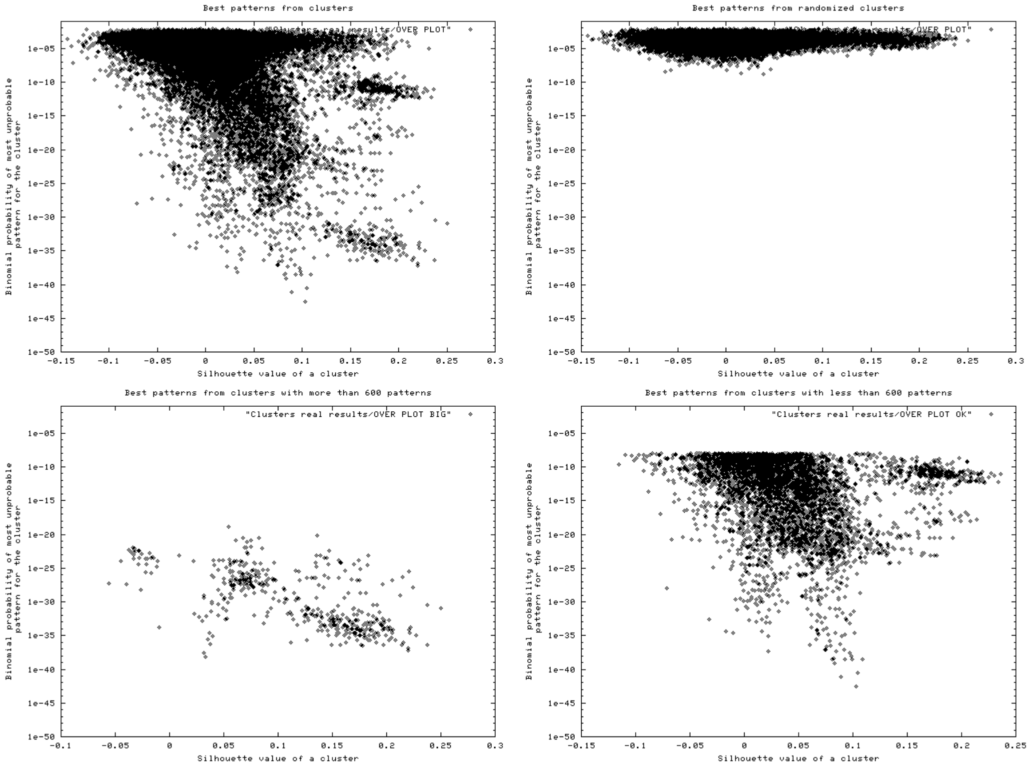

Figure 1: Top left: the silhouette values of the 52 000 clusters (horizontal axis) plotted against the probabilities (vertical axis) of the best substring patterns discovered for each of the respective clusters. Top right: the same for randomized clusters, i.e., for sets of upstream sequences of the same number of randomly chosen genes. If we consider only the patterns that are less probable than the ones discovered from the randomized sets, the plot in the top diagram is stretched along the diagonal, suggesting a correlation between the pattern and the cluster scores. This has to be viewed with caution however, as the clusters are not mutually independent. Bottom left: the silhouette and pattern significance values for significant patterns obtained from clusters that produce more than 600 significant patterns each (these are clusters of highly homologous sequences). Bottom right: the same for patterns obtained from clusters that produce less than 600 significant patterns each. These patterns are candidates for regulatory signals and are subjects of further studies.

exclusively in RAP1, EBF1 binding sites. The first of them (see table below) consisted of 29 patterns, 20 of which had matched RAP1, EBF1 sites. The second one consisted of 15 patterns (see table below) and had 11 matches. Both patterns match different parts of the RAP1, EBF1 site. The alignments are given if Figure 3. The potentially most interesting patterns however are the ones that do not have matches in the known binding sites, and they can be targets for further research.

7 Discussions

As the cluster boundaries from expression studies will not be sharp, the challenge to pattern discovery approaches is the further analysis of all discovered patterns allowing to group together similar patterns and eventually building more general sequence profiles and consensus patterns

```
CLUSTER_0002:  aaTCTTCATGt
CLUSTER_0005:  cgTACCTCTa
CLUSTER_0008:  gACAGCTAc
CLUSTER_0017:  tAT[TAC]GTTAAgc
CLUSTER_0020:  ACTTTATTT
CLUSTER_0021:  [ag]TAACTT[AT]Ca
CLUSTER_0026:  TATCGAG  (SINGLETON)
CLUSTER_0029:  t[ta]CGAATA[AG]aaaa
CLUSTER_0042:  [ta]TGCATGAAc
CLUSTER_0043:  a[TG][GC]GTATAc
CLUSTER_0045:  g[ag][ga][ag][AG][TAG]AT[GA]TG[agt][ga][ag]
CLUSTER_0046:  tag[AG]TAGA[TA]A[ga]aaaa
CLUSTER_0050:  ATCCAAGAg
CLUSTER_0059:  tTTTTCTG[CT][TA]c
```

Figure 2: Consensi of the pattern clusters that do not have matches in SCPD database.

```
-----------------ACCCAGACATCGGGCTTCCAC-
---------------ACACCCAGACATC-----------
---------------ACACCCAGACATC-----------
---------------GAACCCATACACT-----------
---------------ACACCCAGACCGCG----------
---------------GCACCCACACATTT----------
------------GCTAAACCCATGCACAGTGACT-----
-----------------ACCCAGACACGCTCGA------
-------------CTTCACCCTCATAC------------
---------------ACACCCCTTTTCT-----------
---------------GCACCCAGTCTT------------
---------------GCACCCAAACACCTGCATATTTGG
---------------GCACCCAATCACC-----------
---------------ACACCCAGACCTC-----------
---------------AAACCCACACAT------------
--------------TGCACCCATACCTT-----------
--------------AACACCCAAGCACAG----------
-ATCTCTCGCAACG-------------------------
ACCTCCGTACATTC-------------------------
ACACCTGGACACC--------------------------
ACATCCGTACAACGAGAACCCATACATTA----------

     ___              ____
---TCCGTAC---     ACCCATAC--
-CATCCGTAC---     ACCCATACA-
--ATCCGTA----     ACCCATACAT
--ATCCGTACA--     -CCCATAC--
----CCGTAC---     -CCCATACA-
----CCGTACA--     --CCATACAT
---TCCGTACAT-     -CCCATACAT
--ATCCGTACAT-     --CCATACA-
---TCCGTACA--     -AACATAC--
----CCGTACAT-     --ACATACT-
---TCCGTA----     ---GATACT-
--ATCCGTAC---     --AGATACT-
----CCGTACC--
---ACCGTACC--
---ACCGTAC---
--CACCGTAC---
----CCGTACATT
----GCGTAG---
----GCGTAGG--
-----CGTAGG--
-CATCCGTA----
ACATCCGT-----
-CATCCGT-----
```

Figure 3: The upper part of the table shows the alignment of experimentally proven RAP1, EBF1 binding site taken from SCPD database (except the sites `ATGCCCGTGCAC` and `GTCACTAACGACGTGCACCA`, which did not give a good alignment). The alignment on the left, below the site is the cluster 15 of patterns produced automatically by our pattern grouping algorithm (except for the patterns `GTACATT, AACATCCG, TACATCC, ACATCC, ACATCCG`, which do not contain the block CGT). The alignment on the right is the cluster 34 (except for `ACCCA, ACCCAT, ACCCATA`).

Pattern	Probability	Cluster size	Occurrences in cluster	Total occurrences	K
AAAATTTT	2.59075e-43	96	72	830	60
ACGCG	6.41023e-39	96	75	1088	50
ACGCGT	5.23109e-38	94	52	387	40
CCTCGACTAA	5.42764e-38	27	18	23	220
GACGCG	7.88674e-31	86	40	284	38
TTTCGAAACTTACAAAAAT	2.08201e-29	26	14	18	450
TTCTTGTCAAAAAGC	2.08201e-29	26	14	18	325
ACATACTATTGTTAAT	3.80588e-28	22	13	18	280
GATGAGATG	5.59927e-28	68	24	83	84
TGTTTATATTGATGGA	1.8998e-27	24	13	18	220
GATGGATTTCTTGTCAAAA	5.04076e-27	18	12	18	500
TATAAATAGAGC	1.51458e-26	27	13	18	300
GATTTCTTGTCAAA	3.40261e-26	20	12	18	700
GATGGATTTCTTG	3.40261e-26	20	12	18	875
GGTGGCAA	4.17788e-26	40	20	96	180
TTCTTGTCAAAAAGCA	5.09734e-26	29	13	18	250
CGAAACTTACAAA	5.09734e-26	29	13	18	290
GAAACTTACAAAAATAAA	7.9186e-26	21	12	18	650
TTTGTTTATATTG	1.73752e-25	22	12	18	600
ATCAACATACTATTGT	3.62348e-25	23	12	18	375
ATCAACATACTATTGTTA	3.62348e-25	23	12	18	625
GAACGCGCG	4.47204e-25	20	11	13	260
GTTAATTTCGAAAC	7.22797e-25	24	12	18	400
GGTGGCAAAA	3.37381e-24	33	14	31	475
ATCTTTTGTTTATATTGA	7.18849e-24	19	11	18	675
TTTGTTTATATTGATGGA	7.18849e-24	19	11	18	475
GTGGCAAA	1.13567e-23	28	18	137	725
CGAACTGCCAT	1.74392e-23	20	10	10	92
CGAACTGCCATCTC	1.74392e-23	20	10	10	190
CCTCGAACTGCCATCT	1.74392e-23	20	10	10	170

Table 1: The 30 highest scoring patterns discovered in the genome regions upstream from genes of the clusters. Note that the smallest probability of a pattern discovered from the randomized data is 1.74434e-09.

[13].

One way to go, as we see it, is to first use a very restrictive pattern class, for example substring patterns as used here, and then to post-process the output to obtain patterns of a more general class. For example, for a specific cluster, the upstream regions can be mined for unexpectedly frequent substring patterns. Then these patterns can be clustered based on either their mutual similarity or on the locations of their occurrences. The similar patterns or areas of high density of occurrences of unprobable patterns can be analyzed further, for example by a multiple local alignment programs such as MEME [1] or Consensus [13].

References

[1] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21:51–83, 1995.

[2] A. Ben-Dor and Z. Yakhini. Clustering gene expression profiles. In *Proceedings of the Thrid Annual International Conference on Computational Molecular Biology*, pages 33–42. ACM Press, 1999.

[3] A. Brazma, I. Jonassen, I. Eidhammer, and D. Gilbert. Approaches to automatic discovery of patterns in biosequences. *Journal of Computational Biology*, 5(2):277–304, 1998.

[4] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. Predicting gene regulation elements in silico on a genomic scale. *Genome Research*, 8:1202–1215, 1998.

[5] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome wide transcriptional analysis of gene expression of the mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.

[6] S. Chu, J. L. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz. The transcription program of sporulation in budding yeast. *Science*, 282:699–705, 1998.

[7] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.

[8] P. D'haesleer, X. Wen, S. Fuhrman, and R. Somogyi. Mining the gene expression matrix: Inferring gene relationships from large scale gene expression data. In *Information Processing in Cells and Tissues*. Plenum Publishing, 1998.

[9] M. Eisen, P. T. Spellman, D. Botstein, and P. O. Brown. Cluster analysis and display of genome-wide expression patterns. *Proceedings of National Academy of Science USA*, 95:14863–14867, 1998.

[10] B. S. Everitt. *Cluster Analysis*. Edward Arnold, 3rd edition, London, 1993.

[11] K. Frech, K. Quandt, and T. Werner. Software for the analysis of DNA sequence elements of transcription. *Comput. Appl. Biosci.*, 13:89–97, 1997.

[12] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, New York, 1975.

[13] G. Z. Hertz and G. D. Stormo. Identification of consensus patterns in unaligned DNA and protein sequences: a large-deviation statistical basis for penalizing gaps. In *Proc. of Third International Conference on Bioinformatics and Genome Research*. http://www.scri.fsu.edu/mimi/bio94/hertz/hertz.html, 1994.

[14] I. Jonassen. Efficient discovery of conserved patterns using a pattern graph. *Comput. Appl. Biosci.*, 13:509–522, 1997.

[15] Teuvo Kohonen. *Self-Organizing Maps*. Springer, Berlin, Heidelberg, 1997. (Second Extended Edition 1997).

[16] E. M. McCreight. A space–economical suffix tree construction algorithm. *J. ACM*, 23:262–272, 1976.

[17] A. F. Neuwald, J. S. Liu, and C. E. Lawrence. Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. *Protein Science*, 4:1618–1632, 1995.

[18] F. P. Roth, D. Hughes, P. W. Estep, and G. M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantification. *Nature Biotechnology*, 16:939–945, 1998.

[19] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretations and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[20] P. T. Spellman, G. Sherlock, M. Zhang, V. R. Iyer, K. Anders, M. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces Cerevisiae by microarray hybridization. *Molecular Biology of Cell*, 9:3273, 1998.

[21] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of National Academy of Science USA*, 96(6):2907–2912, 1999.

[22] S. Tavazoie, D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22:281–285, 1999.

[23] E. Ukkonen. Constructing suffix trees on–line in linear time. *Information Processing*, 1:484–492, 1992.

[24] J. van Helden, B. André, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology*, 281(5):827–842, 1998.

[25] T.G. Wolfsberg, A.E. Gabrielian, M.J. Campbell, R.J. Cho, J.L. Spouge, and D. Landsman. Candidate regulatory sequence elements for cell cycle-dependent transcription in Saccharomyces cerevisiae. *Genome Research*, 9(8):775–792, August 1999.

Nr	consensus pattern	factors that have matching binding sites
1	tctcaTCTCA[TC][CT][tag]catc	ABF1 ABF1,BAF1 UASPHR
3	cctcGAA[CG]TGCCATCtca	BAS1 BAS1,PHO2 CCBF,SCB,SWI6 HSE,HSTF HSE,HTSF SCB UASH UASPHR XBP1
4	a[ta][CG]CCTA[AT]Aat	MCM1
6	acc[ac]CCCC[CT][CGT][ag]a	MIG1 RAP1 RAP1,EBF1
7	gT[TA][CA]TCCT[CG]g	BAS1 BAS1,PHO2 UASPHR
9	a[ct][at]GTGACA[GTC][cta]t	ADR1 MATalpha1 MATalpha2 MCM1 UASH
10	tt[tc]ACAGT[GT][AT][tc]g	ABF1 ABF1,BAF1 ADR1 BAS1 BAS1,PHO2 GAL4 GCN4 GCN4,GCRE GCRE,GCN4 PHO2 RAP1 RAP1,EBF1
11	[at][ATC]TACACAt	MATalpha2
12	tttGTCACA[GAT]gg	ABF1 ABF1,BAF1 PAE UASH
13	t[gc]ACATT[GC][CT]tg	HSE,HSTF HSE,HTSF PAE RAP1 RAP1,EBF1
14	ata[TC]TGGTTCt	ROX1 URSSGA
15	acaTCCGTAC[acg]tt	HSE,HSTF HSE,HTSF RAP1 RAP1,EBF1
16	a[gca][atc]TAAG[CG][TAG][tga]a	ABF1 ABF1,BAF1 GATA GLN3 MCM1 URS1ERG11
18	t[ct][at][AG]AAGT[AT][TA]c	PRP1 URSPHR
19	gtT[AG]TTA[CT][TG][AG]ca	GRF2 MATalpha2 MCM1 REB1 UASH
22	t[ACT]CGCTTA[AT]	UASGATA
23	gaa[ca][gat][acg][AG]CGCG[cta][gat][ca]gc	ABF1 ABF1,BAF1 CCBF,SCB,SWI6 DAL82 GAL4 HAP1 HAP2 HAP2;HAP3;HAP4 HAP3 HAP4 LEU3 MAL63 MCB PDR1 PDR3 PHO4 RAP1 RAP1,EBF1 REB1 SCB SWI4 SWI6 UASGABA repressor_of_CAR1
24	[ac][at][GT]ACGCcaa	ABF1 ABF1,BAF1
25	GGTCG[CT]Ac	UASPHR URS1ERG11
27	tgtTAACGAATCGTTtaa	GFI,TAF MCM1 TAF
28	ga[at][TC]CGTTTA[ag]g	ABF1 ABF1,BAF1 MAL63 MCM1
30	aA[CAG][AT]GAATCttc	ADR1
31	t[ac][tc][at]CGACT[CA][ca][cg]aa	BAS1 BAS1,PHO2 GAL4 GCN4 GCN4,GCRE GCRE,GCN4 GFI,TAF PHO2 TAF URSSGA
32	tcCACGAA[gc][ta]g	ABF1 ABF1,BAF1 BAS1 BAS1,PHO2 CCBF,SCB,SWI6 GA-BF GFI,TAF HSE,HSTF HSE,HTSF PDR1 PDR3 PHO4 SCB SWI4 SWI6 TAF URS1ERG11
33	c[ga][ctg][ACG]TACG[AT][atc]tat	ABF1 ABF1,BAF1 PHO4 URS1HO
34	aC[CA]CATAC[AT]t	MCM1 RAP1 RAP1,EBF1
35	atat[CT][AG]GCAC[TC][ac]a	GAL4 MCM1 PHO4 RAP1 RAP1,EBF1 URSSGA
36	taGCGCA[GT][ga]cc	ABF1 ABF1,BAF1 ARC CUP2 SWI5 UASPHR repressor_of_CAR1
37	cgGTGGCAA[AC][ag]	ABF1 ABF1,BAF1 HAP2 HAP2;HAP3;HAP4 HAP3 HAP4 RAP1 RAP1,EBF1 UASCAR UASPHR repressor_of_CAR1
38	t[ca][ga][GA]CGGC[TG][GTA][cta]tttt	ABF1 ABF1,BAF1 GAL4 HAP1 LEU3 MCM1 PHO4 QBP RP-A SWI5 UASGABA URS1H URSF URSINO repressor_of_CAR1
39	a[cat][AGC]AGGG[GT][ctg][ac]a	13nt_repeat BUF GAL4 HAP1 IRE MCM1 MIG1 PHO4 RAP1 RAP1,EBF1 RC2;RC1 UAS1ERG11 UAST52,ORE URS1ERG11 URSSGA
40	gcg[ag][at][ga][ac]GATGAG[AC]t[ag][at]g	BUF HAP1 HSE,HSTF HSE,HTSF PQBOX REB1 SWI5 UASH UASPHR
41	aTGGATGCc	MOT3
44	gc[TAG]TATAT[ATC][gat][ag][tg]gg	TATA,TBP
47	gtaTAAATAGAGCtgct	QBP TATA,TBP UIS URS1H URS1HSC82
48	[at]a[ag][TG][AT]GCC[CG][ac][ac]ga	BUF GAL4 GCFAR QBP UME6 URS1H URS1HSC82 repressor_of_CAR1
49	aC[CT]CAAT[AT][tg]t	MATalpha1 MCM1
51	aaacaAAACAAA[AT][ca][ac]aata	GCR1 GCR1,CTBOX MCM1 MSE ROX1 UASPHR
52	tgtGTAAA[TC]ATtt	SFF UAS2CHA URS1ERG11
53	ataaaa[gt][CA][GT]AAAA[GA][cg][gac]aaaag	BAS1,PHO2 CCBF,SCB,SWI6 MAL63 MCM1 MIG1 PHO2 SCB SWI4 SWI5 SWI6 TATA,TBP UASPHR
54	t[gt][TC]GAAAG[AG]Tt	XBP1
55	[at][tac]t[gta][ag]AAAATTTT[tg][tc][at]tt	ABF1 ABF1,BAF1 CSRE DAL82 MAL63 MATalpha2 MCM1 NBF UASH UASINO UIS
56	ga[at][acg][CA]GGAA[AG]T[gt]gaa	GAL4 MCM1 UAS2CHA UASH
57	t[tc][cat][AT][TC]TTC[GA][ACT][ga]t	GAL4 GCR1 GCR1,CTBOX REB1
58	cgg[ct][ctg][gct][ctg]CTTTTT[CTG][TC][atc][tg]cc	ACE1 CUP2 DAL82 GAL4 HSE,HSTF HSE,HTSF LEU3 RAP1 RAP1,EBF1 UASCAR URSSGA
60	t[ta][gta][gtc][TG]TCTA[TG][GTC]a[at][ct]	HSE,HSTF HSE,HTSF ROX1
61	taaat[AT]TTTGTG[ta]ca	MATalpha1 MATalpha2 MCM1 MIG1 UASH
62	t[acg]CTGTG[CT]a[ac]	UASH

Table 2: pattern-clusters Results of matching the discovered patterns against SCPD.

ORF	Gene	Length	Disrupt.	Description
YBL041W ⋆	PRE7	241	lethal	20S proteasome subunit(beta6)
YBR170C	NPL4	580	lethal	nuclear protein localization factor and ER translocation component
YDL126C ⋆	CDC48	835	lethal	microsomal protein of CDC48/PAS1/SEC18 family of ATPases
YDL100C		354		similarity to E.coli arsenical pump-driving ATPase
YDL097C ⋆	RPN6	434	lethal	subunit of the regulatory particle of the proteasome
YDR313C	PIB1	286		phosphatidylinositol(3)-phosphate binding protein
YDR330W		500		similarity to hypothetical S. pombe protein
YDR394W ⋆	RPT3	428	lethal	26S proteasome regulatory subunit
YDR427W ⋆	RPN9	393	viable	subunit of the regulatory particle of the proteasome
YDR510W	SMT3	101	lethal	ubiquitin-like protein
YER012W ⋆	PRE1	198	lethal	20S proteasome subunit C11(beta4)
YFR004W ⋆	RPN11	306	lethal	26S proteasome regulatory subunit
YFR033C	QCR6	147	viable	ubiquinol–cytochrome-c reductase 17K protein
YFR050C ⋆	PRE4	266	lethal	20S proteasome subunit(beta7)
YFR052W ⋆	RPN12	274	lethal	26S proteasome regulatory subunit
YGL048C ⋆	RPT6	405	lethal	26S proteasome regulatory subunit
YGL036W	MTC2	909	viable	Mtf1 Two hybrid Clone 2
YGL011C ⋆	SCL1	252	lethal	20S proteasome subunit YC7ALPHA/Y8 (alpha1)
YGR048W ⋆	UFD1	361	lethal	ubiquitin fusion degradation protein
YGR135W ⋆	PRE9	258	viable	20S proteasome subunit Y13 (alpha3)
YGR253C ⋆	PUP2	260	lethal	20S proteasome subunit(alpha5)
YIL075C ⋆	RPN2	945	lethal	26S proteasome regulatory subunit
YJL102W	MEF2	819		translation elongation factor, mitochondrial
YJL053W	PEP8	379	viable	vacuolar protein sorting/targeting protein
YJL036W		423		weak similarity to Mvp1p
YJL001W ⋆	PRE3	215	lethal	20S proteasome subunit (beta1)
YJR117W	STE24	453	viable	zinc metallo-protease
YKL145W ⋆	RPT1	467	lethal	26S proteasome regulatory subunit
YKL117W	SBA1	216	viable	Hsp90 (Ninety) Associated Co-chaperone
YLR387C		432		similarity to YBR267w
YMR314W ⋆	PRE5	234	lethal	20S proteasome subunit(alpha6)
YOL038W ⋆	PRE6	254		20S proteasome subunit (alpha4)
YOR117W ⋆	RPT5	434	lethal	26S proteasome regulatory subunit
YOR157C ⋆	PUP1	261	lethal	20S proteasome subunit (beta2)
YOR176W	HEM15	393	viable	ferrochelatase precursor
YOR259C ⋆	RPT4	437	lethal	26S proteasome regulatory subunit
YOR317W	FAA1	700	viable	long-chain-fatty-acid–CoA ligase
YOR362C ⋆	PRE10	288	lethal	20S proteasome subunit C1 (alpha7)
YPR103W ⋆	PRE2	287	lethal	20S proteasome subunit (beta5)
YPR108W ⋆	RPN7	429		subunit of the regulatory particle of the proteasome

Table 3: The genes from the `GGTGGCAA`-cluster. The annotations are taken from MIPS database. Note that many of the genes are related to proteasome. The 25 ORFs marked with ⋆ belong to the functional class of "cytoplasmic degradation" containing 93 ORFs in total according to the Functional Catalogue of *Saccharomyces cerevisiae* (MIPS, `http://www.mips.biochem.mpg.de/proj/yeast/catalogues/funcat/`).